CSYE 7245 – Big Data Systems & Intelligent Analytics

Assignment 3

Q1 (5 Points)

Give a brief definition for the following:

1. **Linearity of expectation**

Linearity of expectation is the property that the [expected value](https://brilliant.org/wiki/expected-value-definition/) of the sum of random variables is equal to the sum of their individual expected values, regardless of whether they are independent.

1. **Z-score**

a z-score is the number of [standard deviations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) from the mean a data point is. But more technically it’s a measure of how many standard deviations below or above the population [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean) a [raw score](https://www.statisticshowto.datasciencecentral.com/raw-score/) is. A z-score is also known as a standard score and it can be placed on a [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/) curve.

\_\_

**Zi = (Xi – X )/S**

1. **Chernoff bound**

In [probability theory](https://en.wikipedia.org/wiki/Probability_theory), the Chernoff bound, named after [Herman Chernoff](https://en.wikipedia.org/wiki/Herman_Chernoff) but due to Herman Rubin,[1] gives exponentially decreasing bounds on [tail distributions](https://en.wikipedia.org/wiki/Cumulative_distribution_function#Complementary_cumulative_distribution_function_.28tail_distribution.29) of sums of independent random variables. It is a sharper bound than the known first- or second-moment-based tail bounds such as [Markov's inequality](https://en.wikipedia.org/wiki/Markov%27s_inequality) or [Chebyshev's inequality](https://en.wikipedia.org/wiki/Chebyshev%27s_inequality), which only yield power-law bounds on tail decay. However, the Chernoff bound requires that the variates be independent – a condition that neither Markov's inequality nor Chebyshev's inequality require, although Chebyshev's inequality does require the variates to be pairwise independent.

1. **Monte Carlo algorithm**

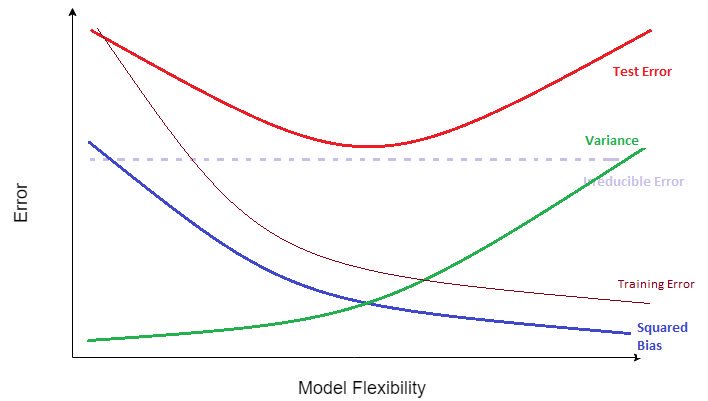
In [computing](https://en.wikipedia.org/wiki/Computing), a Monte Carlo algorithm is a [randomized algorithm](https://en.wikipedia.org/wiki/Randomized_algorithm) whose output may be incorrect with a certain (typically small) [probability](https://en.wikipedia.org/wiki/Probability). Their essential idea is using [randomness](https://en.wikipedia.org/wiki/Randomness) to solve problems that might be deterministic in principle.

Monte Carlo methods vary, but tend to follow a particular pattern:

1. Define a domain of possible inputs
2. Generate inputs randomly from a [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution) over the domain
3. Perform a [deterministic](https://en.wikipedia.org/wiki/Deterministic_algorithm) computation on the inputs
4. Aggregate the results
5. **Bias-variance tradeoff**

If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it’s going to have high variance and low bias. So we need to find the right/good balance without overfitting and under-fitting the data.

This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can’t be more complex and less complex at the same time.



Q2 (5 Points)

Arrange the following functions in increasing order of asymptotic growth:

**0.33n < log n < √n < n2 < n2√n < 5n < 5n5**

Q3 (5 Points)

Master Theorem: For the following recurrence, give an expression for the runtime T(n) if the recurrence can be solved with the Master Theorem. Otherwise, indicate that the Master Theorem does not apply.

1. T(n) = 4T (n/2)+ n

The generic form of the Master theorem –

T(n) = a T ( n / b) + f (n) where a >= 1 , b > 1

From above we can infer :

a = 4 , b = 2 and f( n ) = n

Consider the Case 1 generic form :

f(n) = θ ( nc ) where c < logb a

Since f(n) = n, c = 1.

logb a = log2 4 = 2 which is > c

Hence 1st case of the Master Theorem is followed.

T(n) = θ ( nlogb a) = θ ( n2)

1. T(n) = 4T (n/2)+ n2

T(n) = a T ( n / b) + f (n) where a >= 1 , b > 1

From above we can infer :

a = 4 , b = 2 and f( n ) = n2

Consider the Case 1 generic form :

f(n) = θ ( nc ) where c < logb a

Since f(n) = n2, c = 2 which is not < logb a.

Consider the Case 2 generic form :

f(n) = θ ( nc logk n) where c = logb a

c = 2 and k = 0 and logb a = log2 4 = 2 = c

Hence 2nd case of the Master Theorem is followed.

T(n) = θ ( nlogb a logk+1 n) = θ ( n2 log n)

1. T(n) = 2T (n/2)+ n2 log n

We infer that:

A = 2, b = 2, f(n) = n2log n

Logb a = 1 which is < 2 and so < c

Hence case 1 & 2 are not followed.

T(n) = θ f(n) = θ ( n2 log n)

Case 3 of the master theorem is followed.

Q4 (5 Points)

Stephen Curry hit 77 three-point shots in a row in practice. If his probability of hitting an unguarded three-point shot is 90%, what is the likelihood of Stephen Curry making at least 9 out of 10 three-point shots?

P(hit) = 0.9. P(miss) = 0.1

P(at least 9/10 3 pointers) = P(9 3 pointers) + P(10 3 pointers)

= 10 \* 0.99 \* 0.11 + 1 \* 0.910

= 0.3874 + 0.3487

= 0.7361

Q5 (5 Points)

A booth at the fair has 200 balloons, 5 of which contain $10 and 1 of which contains $20. The rest contain only air. If it costs $1 to randomly break a balloon, what is the expected return of an individual making such an attempt?

P(getting 10$) = 5/200

P(getting 20$) = 1/200

P(getting nothing) = 194/200

Expected return = ((10-1) \* 5/200) + ((20-1) \* 1/200) – 1 \* 194/200

= 0.225 + 0.095 – 0.97

= -0.65

Q6 (5 Points)

Sort the list of integers below using Merge sort. Show your work. Write a recurrence relation for Merge sort.

(22, 13, 26, 1, 12, 27, 33, 15)

Use the Master Theorem to prove the complexity of your recurrence.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Original |  |  |  |  | 22 | 13 | 26 | 1 | 12 | 27 | 33 | 15 |  |  |  |
| / by 2 |  |  |  | 22 | 13 | 26 | 1 |  | 12 | 27 | 33 | 15 |  |  |  |
| / by 2 |  | 22 | 13 |  | 26 | 1 |  |  |  | 12 | 27 |  | 33 | 15 |  |
| / by 2 | 22 |  | 13 |  | 26 |  | 1 |  | 12 |  | 27 |  | 33 |  | 15 |
| Merge 1 |  | 13 | 22 |  |  | 1 | 26 |  |  | 12 | 27 |  |  | 15 | 33 |
| Merge 2 |  |  | 1 | 13 | 22 | 26 |  |  |  |  | 12 | 15 | 27 | 33 |  |
| Final |  |  |  |  | 1 | 12 | 13 | 15 | 22 | 26 | 27 | 33 |  |  |  |

Worst case performance (big-O): O(n log n)

Recurrence relation: T(n) = 2 T(n/2) + n

A = 2, b = 2, f(n) = n

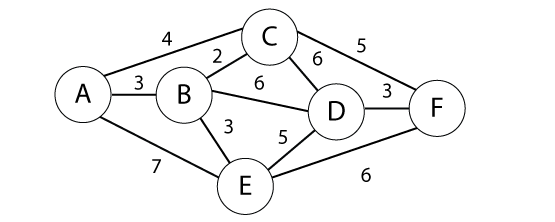
C = 1 = logb a

It follows Case 2 of Masters Theorem where k = 0.

T(n) = θ(n log n)

Q7 (5 Points)

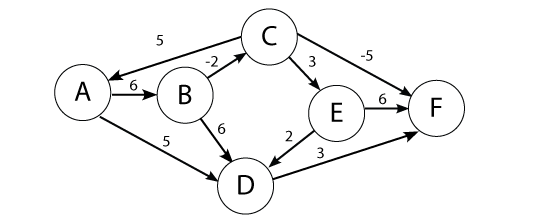
Find shortest path from A to F in the graph below using Dijkstra's algorithm. *Show your steps.*



|  | A | B | C | D | E | F |
| --- | --- | --- | --- | --- | --- | --- |
| A | (0,A) | (3,A) | (4,A) | ∞ | (7,E) | ∞ |
| B |  | (3,A) | (4,A) | (9,B) | (6,B) | ∞ |
| C |  |  | (4,A) | (9,B) | (6,B) | (9,C) |
| E |  |  |  | (9,B) | (6,B) | (9,C) |
| D |  |  |  | (9,B) |  | (9,C) |
| F |  |  |  |  |  | (9,C) |

A -> C -> F (Cost = 9)

Q8 (5 Points)



Use the Bellman-Ford algorithm to find the shortest path from node A to F in the weighted directed graph above. *Show your work.*

|  | A | B | C | D | E | F |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | ∞ | ∞ | ∞ | ∞ | ∞ |
| 1 | 0 | 6 | ∞ | 5 | ∞ | ∞ |
| 2 | 0 | 6 | 4 | 5 | ∞ | 8 |
| 3 | 0 | 6 | 4 | 5 | 7 | -1 |
| 4 | 0 | 6 | 4 | 5 | 7 | -1 |
| 5 | 0 | 6 | 4 | 5 | 7 | -1 |

A -> B -> C -> F (Cost = -1)

Q9 (5 Points)

In a room of 23 people, what is the probability that someone has the same birthday as you?

To figure out the exact probability of finding two people with the same birthday in a given group, it turns out to be easier to ask the opposite question: what is the probability that NO two will share a birthday, i.e., that they will all have different birthdays? With just two people, the probability that they have different birthdays is **364/365, or about .997**.

If a third person joins them, the probability that this new person has a different birthday from those two (i.e., the probability that all three will have different birthdays) is **(364/365) x (363/365), about .992.**

With a fourth person, the probability that all four have different birthdays is

**(364/365) x (363/365) x (362/365), which comes out at around .983.**

And so on. The answers to these multiplications get steadily smaller. When a twenty-third person enters the room, the final fraction that you multiply by is 343/365, and the answer you get drops below .5 for the first time, being approximately **.493**. This is the probability that all 23 people have a different birthday. So, the probability that at least two people share a birthday is

**1 - .493 = 0.507**

Q10 (5 Points)

Not to be done. Waived off by professor.

Q11 (5 Points)

Two linear regression models return t-statistics of 1 and 19 respectively. What is the null hypothesis in this case. Which t-statistic provides more evidence to reject the null hypothesis.

The standard error (SE) is the standard deviation of the sampling distribution of a statistic

A sampling distribution is the probability distribution of a given statistic based on a random sample.

 The dispersion of sample means around the population mean is the standard error. The dispersion of individual observations around the population mean is the standard deviation.

The standard error is an estimate of the standard deviation of the coefficient. It can be thought of as the spread between fitted and actual values.

f(x) = B0 + B1(x) + error

Here, Null Hypothesis is that predicted value f(x) is not dependent on independent value x.

i.e

Is B1 = 0 ?

For a t-statistic high is good, we can think of this if the slope is not small and there is a not much spread between fitted and actual values then we can be confident that the true slope is not 0.

A t-statistic (t value) of greater than 2 in magnitude, corresponds to p-values less than 0.05.

Therefore, if the t-statistics value is greater than 2 we can assume that the null hypothesis is false and f(x) might depend on x.

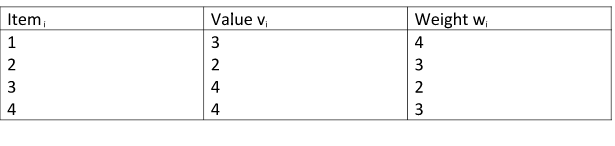
In this example, i.e between T Value of 1 and 19. Considering able explanation, The T value of 19 provides more evidence to reject Null Hypothesis.

Q12 (5 Points)

Given the weights and values of the four items in the table below, select a subset of items with the maximum combined value that will fit in a knapsack with a weight limit, *W,* of 6. Use dynamic programming. *Show your work.*

|  |  |  |
| --- | --- | --- |
| Item i | Value vi | Weight wi |
| 1  2  3  4 | 3  2  4  4 | 4  3  2  3 |

Capacity of knapsack W=6



Knapsack problem:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Val wt | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| (3) 4 | 0 | 0 | 0 | 0 | 3 | 3 | 3 |
| (2) 3 | 0 | 0 | 0 | 2 | 3 | 3 | 3 |
| (4) 2 | 0 | 0 | 4 | 4 | 4 | 6 | 7 |
| (4) 3 | 0 | 0 | 4 | 4 | 4 | 8 | 8 |

Maximum weight = 6

After taking out all the max values and filling the table, we get the above values. Now we consider the value T[3][6] = 8. Since, 8 has not been obtained by the row before, we are considering the item 3 having the value 4. Hence, 3 (4) should be an item in the knapsack